DOCUMENT RESUME

ED 052 667 FL 002 449

AUTHOR Briere, Eugene J.; Brown, Richard H.
TITLE Norming Tests of ESL among Amerindian Children.
PUB DATE 71
NOTE 14p.; Revised version of a paper presented at the Fifth Annual TESOL Convention, New Orleans, La., March 7, 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS American Indian Culture, *American Indians, Educational Diagnosis, Elementary Schools, *Elementary School Students, *English (Second Language), Ethnic Groups, Intensive Language Courses, Language Classification, Language Instruction, Language Proficiency, *Language Tests, Reservations (Indian), Second Language Learning, *Student Evaluation, Testing

ABSTRACT

This paper describes the activities to develop norms for the interpretation of tests designed to indicate proficiency in English for Amerindian children attending grades 3 through 6 in the Bureau of Indian Affairs' schools. The objectives of the test battery are described. The first is to identify the Amerindian child who needs special training in English and to determine the placement in the proper level of intensity of English training. The second purpose is to provide the classroom teacher with specific linguistic information for each child in each language group which could be used as a diagnostic guide for teaching methods or materials. A third objective is to provide a means of assessing the relative merit of various English programs. These objectives require that certain decisions be made which can be classified as placement, diagnostic, and evaluative. The norming group consisted of 7,547 children. Results of this second phase in the evaluation project are discussed. For the companion document see ED034971. (RL)

ED052667

NORMING TESTS OF ESL AMONG AMERINDIAN CHILDREN[1]

by

Eugène J. Brière, University of Southern California

and

Richard H. Brown, University of California at Los Angeles

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

FL 002 449

Beginning in the summer of 1968, the English Language Testing Project[2] started as a research and development program to provide a test of proficiency in English for the Amerindian children attending grades three through six in the Bureau of Indian Affairs' (BIA) schools throughout the United States.[3] The initial phases of the project have been reported previously in Brière (1969 a and b) and in ERIC, May 1970, ED 034 971.

The academic year of 1970-1971 represents the final year for this project. The purpose of this paper is to describe the activities which have taken place to develop norms on which interpretation of test results can be based. To begin with, however, I'd like to describe the test battery very briefly. Three basic types of testing instruments were developed, viz., (1) a written test; (2) a listening comprehension test; and (3) an oral production test.

There are two parallel forms of the written test, each of which contains sixty-two multiple choice items. One type con-

---

[1]This is a revised version of a paper read at the TESOL conference in New Orleans in 1971.

[2]We are grateful for many helpful suggestions made by John Upshur, Chief Testing Officer for the English Language Institute and chief consultant to this project. However, we alone are responsible for any errors or omissions.

[3]This project was funded by the Bureau of Indian Affairs under contract numbers K51C1420092, K51C14200312 and K51C14200382

sists of a question stem which can be answered with one of the choices; e.g., "What does Tommy read in class?" "(a) Yes, he does; (b) Likes books; (c) School books." The second type consists of an incomplete stem which can be completed with one of the choices, e.g., "The _____ in this room is awful." "(a) heat; (b) hot; (c) hotly."

The listening comprehension test consists of aural stimuli, recorded on tape, and three types of multiple choice responses: (a) choosing the correct picture of three which has been described on the tape; (b) identifying factual information which was actually given in a recorded conversation; and (c) using information contained in a recorded conversation in order to infer the correct choice.

The third part of the test battery consists of an oral production test. In this test the student is shown several sets of pictures--each set containing four pictures. Each picture in each of the sets varies slightly from the others along some criterial attribute. The student is then shown a test picture which is identical to one of the four in the set. Two responses are required of the student. First he must point to the picture in the set which matches the test picture. Then he must tell the examiner how that particular picture differs from the others in the set.

In order to aid the classroom teacher in evaluating the children's oral responses, and to standardize evaluation throughout

all of the schools, a correction matrix was designed. On the far left hand side of the matrix is a series of grammatical categories. Each category represents a structure elicited by one of the sets of pictures. Seven require simple sentences and seven require complex sentences in order to describe the picture correctly, e.g., a simple response to one item is "The boys are washing their faces." a response using a complex sentence as in the second half of the oral production test is, "The girl is watching the children read their books." Along the rows opposite each category is a number from one to four which the teacher crosses out if the response is wrong or leaves alone if the answer is right. For each subject tested, the teacher simply adds the column of numbers beneath the child's name, which have not been crossed out.

The reason that the different grammatical categories are assigned numbers ranging from one to four is that, through previous administrations and statistical analyses of the pre tests, it was found that certain categories are more predictive of success or failure on the total tests. The most predictive items are scored four points and so on down to the least predictive items which are scored one point only. All the teacher has to do is listen for one specific grammatical aspect, e.g., plural pronoun agreement, in item four, and allow or disallow the number of points for <u>that category only</u>. In other words, even if part of the child's response is grammatically incorrect, he still receives total credit if the part of the response being evaluated for that particular item is correct. For example, for item number five, where the

category being evaluated is pronoun gender agreement, the response, "The girl are pointing to her mouth." would receive full credit, even though there is an error in number agreement. (So far, one of the most difficult problems we've had is to get teachers to allow full credit for what seems to be a partially incorrect response.)

The objectives of the test battery are threefold. The first is to identify the Amerindian child who needs special training in English versus the child who does not and to determine the placement of the former in the proper level of intensity of training in English. The second purpose is to provide the classroom teacher with specific linguistic information for each child in each language group which could be used as a diagnostic guide for teaching methods or materials. Potentially a third objective is to provide a means of assessing the relative merit of various English programs. These objectives require that certain decisions be made which can be classified as placement, diagnostic and evaluative decisions.

The proper use of any test is to aid in decision making. The decisions made are based on comparisons, e.g., one can compare an individual to another individual, an individual to a group, or a group to another group in terms of their test scores. But in order for these comparisons to be useful in decision making, you have to know certain things about the nature of the comparisons, viz. their stability and the characteristics of the groups they are based on. The determination of group characteristics and statistical indices of stability constitute "norming procedures."

Before any statistical indices of the stability of comparison can be determined, the various factors which can affect comparison stability must be considered. These can be categorized as situational factors, test factors and individual factors.

Situational factors can involve such things as the manner in which a teacher reads the instructions to the class, e.g., relaxed approach or threatening, the characteristics of the room in which the test is administered, e.g., lighting, acoustics, etc., or the introduction of outside interference such as the presence of operating jack-hammers just outside the room. Test factors involve, for example, such things as type of item, content of each item and length of the total test.

The individual factors may be separated again into two classes, transient and stable. Transient factors are those which can change from administration to administration such as the physical and psychological state of a child at the time a test is taken. We assume that stable factors are relatively constant through time in that the rate of change, if any occurs, is extremely slow and include such things as mechanical skill, I. Q., or, in our particular case, ability to use English. With any test what we really want to measure are these stable individual factors and, if we could measure these factors directly, comparisons based on such measurements wouldn't vary much. But in fact, any given test score cannot be assumed to be a direct measure of a stable trait, because situational factors, test factors and transient individual

factors affect the test scores. Therefore, these factors must be taken into account. Those factors over which we have control, must be held constant, and allowances must be made for those we can't control.

Two major steps were taken in the attempt to hold some of the controllable factors constant. A detailed administrative manual was written, and workshops in test administration and scoring were held before the norming phase of the project took place. The directions in the administrative manual consist of three types. The first type provides detailed recommendations covering the procedures which the examiners are to follow in giving the tests. The second type provides simple, step-by-step directions for the written and oral production tests which the examiner is to read to the students. The third, and possibly the most important type, provides detailed directions to the examiner to ensure that he will demonstrate and check the responses to each step of the directions read to the student. In other words, the examiner not only reads the directions and sample items to the student but also performs the specific tasks along <u>with</u> the students and then checks the responses made by each person in the class <u>before</u> the actual test items are presented. The emphasis on the demonstration of each task expected of the student not only ensures the examiner that the student understands and practices the various tasks required on each part of the test battery but also avoids the possibility that the content and grammatical structure of the directions to the student may, in fact,

be far more complex than many of the items on the tests. The last part of the administrative manual contains a number of correct and incorrect responses to the items on the oral production test. The sample responses in the administrative manual were taken from tapes of the responses most frequently made by children in our sample population during the two year pre-testing period of the project. The samples of right and wrong responses given in the manual represent an additional attempt to insure consistent evaluation of the students' oral responses no matter who the examiner is or what group is being tested. The instructions for the listening comprehension tests are tape recorded. Time is allowed between each instruction requiring a response from the student to permit the examiner to demonstrate the desired response. The tape recorded instructions increase the probability of achieving equivalence from situation to situation.

Workshops in test administration and scoring were held in Fairbanks, Alaska and in Flagstaff, Arizona. Through the cooperation of Bureau of Indian Affairs' officials at the agency level and in Washington, D. C., the participants sent to the workshops consisted of administrators, curriculum specialists and classroom teachers. The two workshops lasted three days each and covered the agencies responsible for the teaching of Eskimos, Hopis and Navajos. A third workshop was held at Choctaw Central during the testing of students at Choctaw Central and Connehatta. During the workshops, the participants were given administrative manuals, two forms of

the written test, a tape of the listening comprehension test, the oral production test, answer sheets for the written and listening comprehension tests, and oral production correction matrices. After detailed instructions by the staff from the project, the workshop participants were paired so that one person played the role of an examiner and the other played the role of a student. The roles were then exchanged thereby enabling each participant the opportunity to administer and score the entire battery in order to become completely familiar with all phases of the testing procedure.

The administrative manual and the workshops were designed to hold constant the controllable variables which affect test scores. The attempt to make all test situations exactly the same, can only be partially successful. Further, some aspects can't be controlled. Therefore, some measure must be obtained which will allow quantification of the extent to which the uncontrolled variables affect test scores causing them to depart from direct measurement of the pertinent stable individual factor.

The importance of stability of comparison is the degree to which it affects the relative standings of individuals. This in turn is determined by two other considerations, consistency of score achievement by individuals and the variability or spread of achieved scores. Knowledge of the combined operation of consistency in score achievement and variability of scores allows an estimate of the difference that inconsistency in score achievement makes in an individual's standing with respect to any particular reference group. Since much

is known about the probability of various departures from the average in terms of standard deviation, that is the statistic used to describe variability of assigned scores. A useful statistic for describing consistency of score assignment is the correlation coefficient. It may be based on internal or external consistency. Internal consistency is the degree to which any sub-parts of the test, down to individual items, correlate with each other. Internal consistency measures provide an estimate of what the corresponding external consistency would be. Since these estimates are strongly affected by any series of unanswered items, the estimates must be used with caution. Generally, the actual external consistency is below the internal consistency estimate. (Gulliksen, 1967)

External consistency can be measured in two ways, either by giving the same test again, (test-retest) or by giving parallel forms of the test. The "practice effect" causes trouble with the test-retest method, so that the best method is to give parallel forms. For our written tests we have both internal consistency estimates of reliability, and parallel forms reliability.

The standard deviations of the various tests are based on 5,143 children of the entire sample of 7,547 tested in the fall of 1970. (All children in the third grade were omitted in the computation of the test statistics.) The reliability estimates are based on sub-samples of this total. The internal consistency data for Form A of the written test is based on 291 children, and for Form B on 281. The parallel forms reliability of A and B are based on 502

individuals, 251 of whom had Form A first and 251 of whom had Form B first. Because of the large size of the sample, all the data were coded onto IBM sheets and the calculations were done by computer at the University of Southern California. The mean and standard deviation for the written test for grades four, five and six were 35.6 and 14.2 respectively. The reliability estimates for Form A were KR-20 .95 and Form B KR-20 .96. The parallel forms reliability is .89 from Form A to Form B and .90 from Form B to Form A. As might be expected, there was a noticeable practice effect from A to B and from B to A. Thus the mean score for Form A is 34.4 when given first, and 36.3 when given after Form B. Similarly the mean score for Form B is 33.5 when given first, and 36.1 when given after Form A.

The statistic combining the measure of variability of scores and the measure of consistency in score achievement is the standard error of measurement. It reflects the relative error introduced by using obtained scores to make comparisons instead of using direct measures of the trait, or "true" scores. For our written tests, using the parallel forms reliability, the SE_m is 4.6 raw score units. This means that 68 percent of the time, or two times out of three, a student's "true" score will fall in a range that is 4.6 points above or below his obtained score.

In addition to knowing the stability of comparison, as reflected by the standard error of measurement, for comparisons based on any particular test in the battery, it is useful to know

the degree to which scores on the various different kinds of tests in the battery covary. A measure of the extent to which any individual's score on one part of the test can be used to predict his score on another part is an indication of the degree to which the different parts of the test battery actually assess different aspects of ability to use English. Using the correlation coefficient, the degree of correspondance of scores between different parts of the test are written and listening .51, written and oral production .50, listening and oral production .43. The magnitudes of these inter-correlations indicate that only about 10 to 15 percent of the variation of the scores on one part can be accounted for by knowing scores on any other part. Thus it can be concluded that the different parts of the test are in fact assessing different aspects of ability to use English.

As noted above, for comparisons to be useful, not only must there be information about the stability of comparison, there must also be information about the characteristics of the groups the comparisons are based on. Properly speaking, the statistics noted above are valid only for groups similar to the groups they were developed from.

The norming group consisted of 7,547 children. This group is comprised of every Amerindian child in grades three, four, five, and six in the schools selected for the norming procedures. In addition, in order to control the "time" variable, the children were tested on, or immediately following, the twentieth day of

instruction. (In determining precisely what the twentieth day of instruction was in each of the schools, holidays were not counted.) There were 1,070 Eskimos from 23 schools in Alaska, 5,898 Navajos from 26 schools in the Southwest, 314 Hopis from 3 schools in Arizona and 265 Choctaws from 2 schools in Mississippi. Excluding the sample of 502 who provided data on the correlation of Forms A and B, there were 1,802 third graders, 1,894 fourth graders, 1,648 fifth graders and 1,633 sixth graders. The correlation of grade with age is about .30, indicating that the primary criterion in grading is age.

Of the 54 schools, fifteen had enrollments of over 600, twelve from 300 to 599, eight from 150 to 299, seven from 75 to 149, and twelve of less than 75. Nine schools were rated as having easy access, eighteen were rated as having difficult access, and twenty-seven as extremely difficult access or quite remote. We included the factors of accessibility as one of the variables affecting language behavior of the children on the basis of Spolsky's 1970 study.

By June 1971, we expect to have testing instruments, an administrative manual, an interpretive manual and norms established for all children in the intermediate grades in BIA schools. The question of the practicality of using these tests for native speakers of other languages, say Spanish, and establishing norms for these other groups is an inviting path to follow in future research.

References Cited

Brière, E. J., "ESL Testing on the Navajo Reservation," TESOL Quarterly, March, 1969 (a).

_____, "Testing ESL Skills Among American Indian Children," 20th Georgetown Roundtable on Linguistics, Vol. XXII, 1969 (b).

Gulliksen, Harold, Theory of Mental Tests. New York: John Wiley & Sons, 1967.

Spolsky, Bernard, "Navajo Language Maintenance: Six-Year-Olds in 1969," Progress Report No. 5, March, 1970.